Selecting and Using Computer-Based Language Tests (CBLTs) to Assess Language Proficiency: Guidelines for Educators

Kim MacDonald, Jean Nielsen, and Lisa Lai

With the growing demand for and use of computer-based language tests (CBLTs) comes the need for clear guidelines to help educators as they attempt to select appropriate tests to assess their students with respect to their second- and foreign-language (L2/FL) teaching-learning goals. The purpose of this article is to provide guidelines to educators who are seeking appropriate CBLTs to assess language proficiency in the classroom. We begin with an introduction that includes a brief word about our intended audience, our rationale for creating this set of guidelines, and the development procedure used. We continue with a discussion of some relevant assessment criteria and conclude with a few summary remarks. Finally, we present a "CBLT Selection and Use: Guidelines Summary Table" composed of questions to ask and an accompanying checklist.

De l'augmentation dans la demande de tests langagiers informatisés (computer-based language tests - CBLT) découle le besoin d'établir des principes directeurs pour aider les enseignants à faire le meilleur choix de tests d'après leurs buts précis quant à l'enseignement/l'apprentissage en langue seconde ou langue étrangère. L'objectif de cet article est de fournir des lignes directrices aux enseignants à la recherche de test langagiers informatisés pour évaluer les compétences langagières de leurs élèves. L'introduction comprend un aperçu de la clientèle cible, une explication des raisons qui nous ont incités à établir ces principes directeurs et une description du processus employé pendant le développement. Une discussion de certains critères d'évaluation est suivie de quelques remarques récapitulatives. Le tableau qui vient ensuite résumer les principes directeurs proposés et l'emploi des tests informatisés comprend des questions qui aident à diriger les choix, et une liste de vérification.

Introduction

Intended Audience

This set of criteria and assessment procedures is intended to serve as a guideline for educators seeking an appropriate computer-based language test (CBLT) to assess language proficiency in the classroom. Although it is normally universities and colleges that have the funds to invest in and the

resources to implement language labs or multimedia centers, this does not suggest that the guidelines provided are exclusive to CBLT decision-making by postsecondary educators. In fact, this set of guidelines may be usefully implemented by educators at any level and for many types of assessment purposes (such as achievement or placement).

Rationale

As technology improves, educators are increasingly turning to CBLTs as fast, reliable, and more accurate measures of student learning.[1] CBLTs provide advantages in scoring, adaptation to student ability, reliability of test items based on item response theory and expert system models, and greater student motivation (Lindblad, 1981; Larson, 1985; Luk, 1991; Thelwall, 2000). With the growing demand for and use of CBLTs comes the need for clear guidelines to help educators as they attempt to select an appropriate test for their individual second-language teaching-learning purposes (Chapelle, 2001). That this need has not yet been adequately met became clear during our own experiences conducting reviews of two published and widely used computerized language tests, the Computer-Based TOEFL-CBT (Lai & MacDonald, 2001) and the Levels of English Proficiency-LOEP (Nielsen, 2004). While conducting these reviews, we were hampered by the lack of guidelines in the field, which was compounded by our own relative inexperience in CBLT evaluation. With our combined computer and language teaching experience, if we met with such difficulties, how much harder would it be for the average educator with little experience in the field? Our goal for the following CBLT guidelines and procedures is twofold; we would like to give educators appropriate criteria with which to assess a particular CBLT, and we would like to save them time as they do it.

Development Procedures

Using both personal experience with computer-based testing and information from assessment literature (Brown, 1997; Cohen, 1994; Cumming, 1996; Dunkel, 1999; Thelwall, 2000), we identified the relevant criteria to consider in CBLT evaluation (discussed below), and categorized them into six sections of issues. Our objective is to help educators better understand these issues and determine how they address the assessment of a given CBLT. Next, we apply the criteria in a "CBLT Selection and Use: Guidelines Summary Table." The table consists of questions to ask and an accompanying checklist. The questions serve as a means of prioritizing criteria, and the three-column checklist consists of *yes*, *no*, and *unsure*. We would stress that this checklist is in no way meant to make the decision for the educator. Because everyone will have different needs and circumstances when first approaching a CBLT, it would be impractical to suggest that a predetermined number of checks in the *yes* column automatically means that the particular test under review is

acceptable. Instead, the purpose of the checklist is to help educators organize their thoughts and keep track of the issues they need to review before making their final decision. For example, when the answer to a particular question is *no*, educators are advised to stop and determine how important that particular criterion is to their assessment purposes. A case in point would be a low-stakes CBLT, which may not require the lengthy development and piloting that a high-stakes test would require. In this situation, an answer *No* to a question about extensive piloting would not necessarily be a problem. Clearly educators need to think about what they are testing before they go through the checklist because their assessment goals will affect their responses to the questions. It is our hope, however, that the criteria we provide will guide them as quickly and efficiently as possible in the thinking process.

Discussion

Discussion of Assessment Criteria

Assessment criteria can be grouped into five categories: Basic Assessment Issues, Scoring Issues, Design Issues, Administrative Issues, and Technical Issues (hardware and software). Relevant criteria to consider include:

Basic assessment issues

The purpose of the assessment. For example, the American Council on Education states that CBLTs can be used for a wide variety of purposes, each purpose served by a specific CBLT whose major function might be mastery, proficiency, achievement, or placement (Dunkel, 1999). At the same time, it is important to ensure a balance between test task and the target language skill to be assessed. According to Green et al. (1995), it is significant that CBLTs provide adequate assessment for the entire range of abilities from low to high as represented by the examinee population.

Reliability. Reliability refers to the precision and consistency of scores derived from the CBLT. Factors that affect reliability are general, situational, and individual (Dunkel, 1999).

Validity. Validity concerns the degree to which the CBLT actually measures what it purports to measure. It is often divided into content, construct, criterion, concurrent, and consequential validity (Dunkel, 1999). As with any second-language assessment instrument, validity is of utmost concern (Bachman, 2000; Cumming, 1996). However, new assessment mediums require new considerations. Users should have an understanding of the inferences made from scores on computer- or Web-based language tests.

Access and understanding of published specification tables. It is important that there be access to published specification tables in order to assure understanding of test scores, the exact focus of the test, its content and type of

language tasks, as well as the mastery-level scale with corresponding cut points (Dunkel, 1999).

Scoring issues

The psychometric model used. Educators should understand the assumptions underlying the psychometric model (Item Response Theory and others) used in the construction of the CBLT. Such understanding helps educators to select the test that has the best fit with the purpose of the assessment and the resources available for trailing.

Scoring methods. Depending on the purpose of the CBLT and the type of information to be interpreted from the scores, several scoring methods can be used. They include raw scores (a simple sum of the number of correct items), weighted raw scores (in which some items count more than others), and scaled scores (e.g., the TOEFL, which is equated across forms), correction for guessing, and so forth (Brown, 1997).

Item omission. On CBLTs this issue can be problematic for several reasons; if omitted items are not scored, test manipulation can occur, which results in undeservedly high scores. Item omission can result in item bank exposure and ensuing test security issues; omission patterns could be linked to language proficiency as measured on such a test and would become a source of measurement error (Brown, 1997).

Cut points or standard setting . Several theories are proposed to determine cut points or standard setting, and each performs a specific purpose depending on the objective of the CBLT. Some include the sequential probability ratio test (SPRT) for making pass-fail decisions in adaptive and other related tests and the Bayesian decision theory for making mastery/non-mastery decisions on a computerized sequential mastery test (Brown, 1997).

Design issues

Item banking. It is important that respondent performance on one test item be consistent with performance on another item. Piloting of tests is especially important in computer-adaptive tests because new tests are created every time (Brown, 1997). Consistency can be aided through test item control. The storage and categorization of test items in a CBLT can add a dimension to the traditional card system, for example; this is important in the number of test subsets that can be created (Brown, 1997).

Security. Given the nature of a CBLT, special attention must be paid to security.

The type, length, and timing of the exam. These criteria need to be appropriate to the aims of the test. With adaptive tests, the computer program selects test items according to how the test-taker has done on the previous test item. Fixed-length tests have a predetermined number of items, whereas open tests will have as many items as are needed to determine the proficiency of the test-taker. However, there is a possible bias in open tests toward students

who end up with short test versions. According to Brown (1997), this is an area still to be investigated more fully. Self-pacing tests allow the test-takers to work at their own pace, whereas timed tests give test-takers a set amount of time to finish. Dunkel (1999) suggests that the "speed of examinee responses can actually be used as additional information in assessing proficiency, if desired and warranted" (p. 79).

Administrative issues

Transient factors. The physical or psychological health of the respondents (Cohen, 1994; Dunkel, 1999).

Stable factors. This refers to the respondent's experience, either with similar computer-based language tests or familiarity with computers in general. Cohen (1994) reports a study showing that CBLT results are indeed influenced by past expertise. Brown (1997) points to similar studies with computerized TOEFL tests, but also refers to studies that indicated that "after students participate in a computer-based testing tutorial, there is no meaningful relationship between computer familiarity and individuals' TOEFL scores" (p. 47).

Learner characteristics. Technological advances in education have meant that even learners who are not comfortable with using computers are expected to use them. In this case, it may not be a matter of being unaccustomed to computers, but of being reluctant to or afraid of using them (computer anxiety or technophobia).

Learner disabilities. All tests, whether traditional or computer-based, need to take into consideration any special needs a student may have. Computer-based language tests may have some extra challenges.

Overall ease of use. In a CBLT, layout and presentation can make all the difference in how easily the test-taker can respond to the test questions.

Technical issues (hardware)

Computer hardware status. Acquiring new computer equipment is extremely costly, but as technology advances, equipment can often become quickly outdated. There is no point in choosing a test that has hardware requirements that the target educational institution cannot support. Something educators need to think about is whether they should be evaluating only CBLTs that suit existing equipment, or choosing better CBLTs that will incur additional upgrading costs.

Back-up: in case of electrical failures, server problems, and so forth.

Technical issues (software)

Delivery system. Educators interested in using CBLTs must take into consideration the delivery system of the test in question. Delivery systems include: stand-alone (a system in which each station is started, stores data and provides individual results independently of the others); networked (a

system that is started, stores data, and provides results from a local server); and Web-based (a system that is accessed, stores data, and provides results via the Internet).

User-friendly interface and display. This refers to the degree of ease, simplicity, and comfort with which the users are able to interact with the CBLT. For the examinees, this characteristic of the CBLT can influence their ability to interact with the CBLT in the most comfortable manner possible, thus lessening examinee anxiety and increasing the possibility of positive and reliable test results. For the educator, it can influence the accuracy of the data retrieval and analysis process, record-keeping, and maintenance procedures.

Support. Educators should be aware of essential and available support needed to ensure efficient use and proper maintenance of the CBLT. Inadequately supported and maintained software will quickly negate any positive attributes of the CBLT, which in turn will negatively influence the CBLT reputation.

Conclusion

The guidelines presented in this article are not exhaustive. In fact they will quickly become out of date as new technological advances move the bar higher in terms of what kinds of assessment can be done, how, and when. This does not mean our attempts to create guidelines have been futile. Instead, educators can use the criteria and procedures presented here as a useful starting point. Even though the parameters keep shifting, we believe that CBLTs have the potential to reduce the burden of assessment while helping us to keep our eyes on the main goal: valid and reliable testing of students' language competence and proficiency. We developed this set of procedures to show the areas that need to be considered when choosing a CBLT, but we acknowledge that our checklist has not yet been piloted. Still, it is our hope that our assessment tool has at least gathered in one easy-to-use document the main issues and questions to reflect on when choosing a CBLT. With more personalized testing (computer-adaptive and intelligent systems), CBLTs can be used to promote deeper and more effective learning in a range of skills, knowledge, and understanding. Before this happens, however, L2/FL educators will need to devote more time to learning about how and when best to implement innovative CBLT assessment methods. Unfortunately, as we have learned from this project, there is still a long way to go in the field of CBLT assessment, and many of the issues have yet to be adequately researched.

CBLT Selection and Use: Guidelines Summary Table

Basic Assessment Issues

Questions to ask	*Yes*	*No*	*Unsure*
1. Does the CBLT clearly state the purpose of the test as formulated by the developers?			
2. Is the CBLT designed in such a way as to provide adequate assessment for the entire range of ability, from low to high, as represented in the examinee population? (e.g., in the case of listening skills, do the items in the pool cover low to high listening ability levels, as well as a variety of listening tasks?).			
3. General			
(a) Does the CBLT item pool sample the stated test objectives?			
(b) Does it lessen the ambiguity in interpreting the test items?			
4. Situational: Is the testing environment suitable for CBLTs?			
5. Individual: Is the tension as a result of taking the CBLT minimized as much as possible? (See Administrative Issues)			
6. Do research reports and other documents provided by the developer show evidence of the following types of test validity?			
(a) Content validity: Does the CBLT adequately sample the content of the subject matter?			
(b) Construct validity: Do the CBLT scores permit accurate inferences about the underlying traits being measured?			
(c) Criterion validity: Is there evidence that examinees' performance on specific sets of objectives on the CBLT is similar to their performance on other instruments?			
(d) Concurrent validity: Is there evidence that examinees' performance on the CBLT parallels that on other CBLTs?			
(e) Consequential validity: Are there potential consequences of the CBLT? (Some factors to consider: Are the rights of the examinee considered? Are the institution's responsibilities outlined? Is the public interest considered? Is the value system that informs the use of the test clearly stated?)			
7. Does a currently published specification table accompany the CBLT?			

Scoring Issues

Questions to ask	*Yes*	*No*	*Unsure*
1. Is there a link between the psychometric model framing the CBLT and the construct being measured?			
2. How is the CBLT scored?			
3. Have ample explanations concerning the scoring methods in nontechnical terms been provided so educators can understand and interpret the results?			
4. Are item omissions permitted?			
5. If so, is there a procedure for dealing with CBLT item omission? (e.g., Are missed items presumed to be correct or incorrect?			
6. If incorrect, is the effect of such a wrong item on the estimation of the items that should follow considered?			
7. Does the cut-point method help with decision making with respect to the CBLT?			

Design Issues

Questions to ask	*Yes*	*No*	*Unsure*
1. Item tasks:			
(a) Are the items appropriate to the task? (e.g., If for reading comprehension, is the length of the reading passage acceptable and feasible given the time limitations, scrolling demands etc?)			
(b) Is there an ample range of tasks, and selected at random?			
2. Item exposure:			
(a) Is the size of the item pool large enough to prevent memorization on test repeats?			
3. Piloting:			
(a) Has this test been piloted?			
(b) If not, is there evidence of adequate sampling variation, and item checking?			
4. Item omission:			
(a) Can students omit an item, or do they have the ability to go back and change an answer?			
(b) If yes, is this appropriate to the test goals?			
5. Are there safeguards in place to prevent competent hackers from breaking in and obtaining the test items?			
6. Are the people who have access to the scores, the ones who SHOULD have access?			
7. Web-based tests: Is the internet site secure?			
8. Adaptive tests: Is there an appropriate stopping rule?			

9. Do you agree with the criteria on which this decision was based?
10. Is it a fixed-length test?
11. Is it an open test?
12. Is it self-pacing?
13. Is it timed?

Administrative Issues

Questions to ask	*Yes*	*No*	*Unsure*
1. Does the CBLT administration process provide features that minimize and account for the influence of negative performance factors such an examinee illness, fatigue, apathy, poor motivation, or personal problems?			
2. Does the CBLT require special computer expertise?			
3. Will students who are accustomed to using computers have a significant advantage over those who aren't?			
4. Will it be difficult for students who are new to this format to become accustomed to this particular CBLT?			
5. Does the examinee have the ability to review instructions, if needed?			
6. Does the CBLT adequately explain to examinees the CBT system, how it works, and how the results are scored?			
7. Does the CBLT administration process positively contribute to examinee comfort level? For instance, does it provide a means to reduce computer anxiety in respondents, either by practice language tests, extra computer work, support, and so forth?			
8. Is there an equivalent pen-and-paper backup exam available?			
9. Can special-needs examinees (e.g., sight-challenged) still take the CBLT using special software/hardware—and is the information about this provided?			
10. Does the CBLT require extensive use of			
(a) Keyboard skills?			
(b) A computer mouse?			
(c) Ability to distinguish between a multitude of windows?			
11. If yes, is this appropriate to the needs of the test?			
12. Does the CBLT administration adequately provide orientation?			

Technical Issues (Hardware)

Questions to ask	*Yes*	*No*	*Unsure*
1. Age of computers: Are breakdown and parts replacement likely to be issues?			
2. Storage capacity of computers: Can they handle multimedia tasks using audiofiles, graphics, video clips, and so forth?			
3. Appropriate peripherals: Will there be a need for headphones? Microphones? Are they available?			
4. Is there a DOS back-up of the exam?			
5. Is there a pen-and-paper back-up?			
6. Is there any provision made for on-line failures/server problems?			

Technical Issues (Software)

Questions to ask	*Yes*	*No*	*Unsure*
1. Is the L2 CBT a stand-alone?			
2. Is the L2 CBT networked?			
3. Is the L2 CBT Web-based?			
4. Is there congruence between the hardware available and the delivery system of the required CBLT?			
5. Does the CBLT software contain quality animation, graphics, and full-motion video?			
6. Does it have adequate input devices?			
7. Does the software contain menus for the examinee?			
8. Does it have adequate student control?			
9. Does it allow for record keeping, data analysis and generation of new items?			
10. Is the software secure and accessible?			
11. Is the CBLT software adequately maintained either by in-house or online support, manuals, and so forth?			

Note

[1]CBLT history dates back nearly two decades. In 1981, because of the difficulty and time involved in scoring open-ended tests, a system of computer programs for the feeding in of tests and their statistical processing (called PRINS) was developed for the scoring function. In practice, computers were used mainly to act as a medium for the test and to score tests. They were not used to deliver the test in any intelligent manner until the Computerized Adaptive Spanish Placement Test was first administered in Brigham Young University (Larson, 1985). In this test, the computer was able to adapt the test questions to suit individual examinees. Using a large item bank, the test presented examinees with items that estimated their range of abilities, and gave them immediate feedback as to their probable class placement. Further innovations came in the late 1980s when the use of computer quizzes based on a set of lessons corresponding to individual textbook chapters (assessing student learning achievements in French) was reported in a study conducted in Southwest Texas State University (Fischer, 1989).

Meanwhile, expert systems involving intelligent selection of items started to be devised to assist adaptive testing systems (Frick, 1989). In summary, CBLTs used at the university level in the past two decades have developed into two main streams: computer-assisted language testing and computer-adaptive language testing, facilitated by item banking or expert systems (Fisher, 1989; Frick, 1989; Brown, 1997)

The Authors

Kim MacDonald has a BSc and an MEd in curriculum studies from St. Francis Xavier University and a BA in French and a BEd from Saint-Anne's University. At present she is an doctoral candidate specializing in second-language education in the Curriculum, Teaching and Learning Department, OISE, University of Toronto. Kim is Coordinator of the Multimedia Language Centre, St. Francis Xavier University in Antigonish, Nova Scotia (http://www.stfx.ca/pinstitutes/). E-mail: kmacdona@stfx.ca or kamacdonald@oise.utoronto.ca

Jean Nielsen has an MA in English from York University and an MEd in second language education from OISE, University of Toronto. Jean is a professor in the English Language Institute at Seneca College in Toronto (http://www.senecac.on.ca). E-mail: jean.nielsen@senecac.on.ca

Lisa Lai has a doctorate in computer applications from the Curriculum, Teaching and Learning Department, OISE, University of Toronto. Lisa is an assistant professor in the English Division, Tatung University in Taipei,Taiwan (http://www.ttu.edu.tw). E-mail: lisahlai@yahoo.com or lisalai@ttu.edu.tw

References

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Brown, J.D. (1997). Computers in language testing: Present research and some future directions. *Language Learning and Technology, 1*(1), 44-59.

Chapelle, C.A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research.* Cambridge, UK: Cambridge University Press.

Cohen, A.D. (1994). *Assessing language ability in the classroom* (2nd Ed.). Boston, MA: Heinle & Heinle.

Cumming, A. (1996). The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 1-14). Clevedon, UK: Multilingual Matters.

Dunkel, P.A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology, 2*(2), 77-93.

Fischer, B. (1989). Instructional computing in French: The student view. *Foreign Language Annals, 22*(1), 79-90.

Frick, T.W. (1989). *EXSPRT: An expert systems approach to computer-based adaptive testing.* (ERIC Document Reproduction Services No. ED 307319)

Green, B., Kingsbury, G., Lloyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J., Zara, T., & Schwartz, J. (1995). *Guidelines for computerized-adaptive test development and use in education.* Washington, DC: American Council on Education Credit by Examination Program.

Lai, L., & MacDonald, K. (2001). *A review of the Computer-Based TOEFL (CBT): Evaluating technical innovations of the CBT.* Unpublished manuscript.

Larson, J.W. (1985). *Computerized adaptive Spanish placement test.* (ERIC Document Reproduction Services No. ED 355772)

Lindblad, T. (1981). *Computer-based analysis of open-ended foreign-language tests items.* (ERIC Document Reproduction Services No. ED 217732)

Luk, H. (1991). *An empirical comparison of an expert systems approach and an IRT approach to computer-based adaptive mastery testing*. (ERIC Document Reproduction Services No. ED 334210)

Nielsen, J. (2004). *Levels of English Proficiency (LOEP): A computer-adaptive language test*. Unpublished manuscript.

Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computers and Education, 34*, 37-49.